

AD-A181 945

A PARALLEL MODEL OF THE KINETIC DEPTH EFFECT USING  
LOCAL COMPUTATIONS(U) NEW YORK UNIV NY COURANT INST OF  
MATHEMATICAL SCIENCES M S LANDY JUN 86 TR-86-2

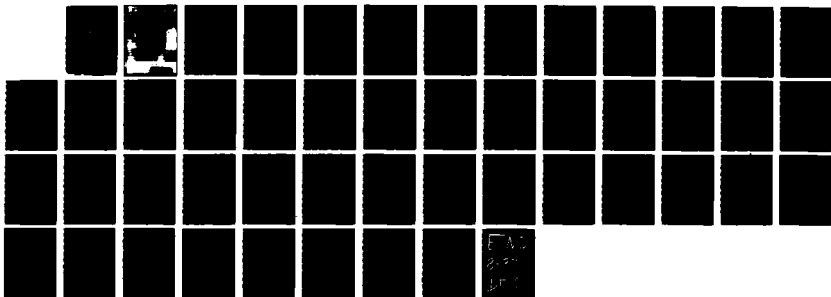
1/1

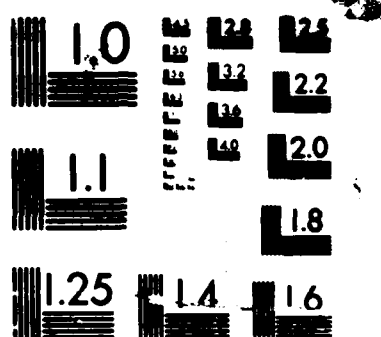
UNCLASSIFIED

N00014-85-K-0077

F/G 23/3

NL





MICROCOPY RESOLUTION TEST CHART

FILE COPY

11

# Computer Science Department

AD-A181 945

## TECHNICAL REPORT

A Parallel Model of the Kinetic Depth Effect  
Using Local Computations

Michael J. Tarr

Computer Science Department, NYU

Modeling Processes in Perception and Cognition

Technical Report #2

June 1987

Psychology Department

### NEW YORK UNIVERSITY

DISTRIBUTION STATEMENT A

Approved for public release,  
Distribution Unlimited

DTIC

LECTURE

JUN 22 1987

Department of Computer Science  
Courant Institute of Mathematical Sciences  
251 MERCER STREET, NEW YORK, N.Y. 10012

## **A Parallel Model of the Kinetic Depth Effect Using Local Computations**

by

*Michael S. Landy* †  
Psychology Department, NYU

---

Mathematical Studies in Perception and Cognition  
Technical Report 86-2  
June, 1986

†Psychology Department  
New York University  
6 Washington Place, Office 961  
New York, NY 10003 USA  
landy@nyu.arpa

To Appear in *Journal of the Optical Society of America A*, 1987

The work described in this paper was supported in part by a grant from the Office of Naval Research, Grant N00014-85-K-0077, Work Unit NR 4007006. Thanks are given to George Sperling, who suggested the line of research and the approach. Thanks are also due Barbara Doshier, Bob Hummel, Charles Cubb, and Mark Perkins for many helpful discussions, suggestions, and encouragement.

**DISTRIBUTION STATEMENT A**

Approved for public release  
Distribution Unlimited

Mathematical Studies in Perception and Cognition

86 - 2

**A Parallel Model of the Kinetic Depth Effect Using Local Computations**

*Michael S. Landy*

Psychology Department  
New York University

Running Head: KDE Model

Keywords: KDE, Kinetic Depth Effect, Relaxation Labeling, Structure From Motion,  
Additive Cues

In Press: *Journal of the Optical Society of America A*, 1987.

Address correspondence to:

Michael S. Landy  
Psychology Department  
New York University  
6 Washington Place, Rm. 961  
New York, New York 10003

Accession For	
NTIS CRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By <i>ltr. on file</i>	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
<i>A-1</i>	



2 -

## Abstract

This paper defines a new model for the Kinetic Depth Effect for multi-dot stimuli. The calculation is performed in a cooperative-competitive network, described as a relaxation labeling process (RLP). The process involves a local iterative computation to best meet the constraints indicated by image cues to depth. Given a constraint that prefers inter-dot distances in 3D to remain constant (local rigidity), the model becomes a local parallel computation of the Ullman incremental rigidity scheme. Several simulations of the model are described, including some where additional cues are combined with the changing dot position cue.

*Keywords: Motion vision; information extraction.*

## 1. Introduction

The human visual system extracts information concerning the 3-dimensional structure of objects using a large variety of cues. These cues derive from a number of the factors which affect the stream of retinal images on our two eyes<sup>1</sup>, including the geometry of viewing from two positions (binocular stereopsis), the geometry of projection (e.g. shape from perspective drawings, texture gradients, and if the viewer or object moves, motion parallax and the kinetic depth effect or KDE), the minutiae of visual optics (cues from accommodation, focus, and chromatic dispersion), and the characteristics of the light-carrying medium (distant objects appear hazy and bluish). These cues may act either alone or in concert to result in the 3-dimensional percept. For example, the cue to depth from relative motion of objects (the KDE) can be very effective even in the absence of other cues to depth<sup>2</sup>, but additional cues may be used together to control the particular percept chosen among several ambiguous possibilities<sup>3</sup>.

In the case of the kinetic depth effect, there has been a great deal of work concerning the determinants of the effect. Using displays consisting of a number of luminous points in a rigid three-dimensional configuration rotating about a fixed axis (Fig. 1), the strength of the perceived depth impression is controlled to varying degrees by the number of points, the speed of rotation, the presence and degree of polar projection, occlusion of farther points by nearer point-containing "objects", etc.<sup>4</sup>.

-----  
Insert Figure 1 about here  
-----

There have been several recent attempts to model the computation of depth values from relative motion<sup>5</sup>. If a given set of points has been viewed in motion over time, and one assumes that these points were in a rigid 3D configuration (the "rigidity assumption"), it is possible that the geometry is sufficient to specify the unknown depth values. This approach to the problem has led to several "n views of m points" results<sup>6</sup>. These results take as their input the 2D image coordinates of the points for a sequence of frames. General assumptions are made about the positions of the points relative to one another and (perhaps) the axis of motion, allowing the depth values to be derived. The geometry of parallel projection is such that these values can only be known up to an additive constant (i.e. the distance can only be known in relative terms), and up to a possible reflection about the image plane (the reversal of depth and motion direction often seen in these multi-dot displays; such reversals are also visible in static displays such as the Necker cube, Fig. 11-a).

A second approach is to use more measurements at each object point. For example, in addition to measuring object position in the image plane, one might also compute the derivatives of these measurements (i.e. the velocity vector in the image plane). This results in an optical flow map across the image. There are several models which utilize this flow in addition to the positions<sup>7</sup>.

A recent model by Ullman<sup>8</sup> uses a very elegant scheme to compute depth values. Since it is closely related to the model discussed in this paper, we will describe it in some detail. Its input consists of a sequence of point positions and correspondences between the points in successive frames. A generalization of the rigidity assumption is used. Rather than assuming that the object is totally rigid, it assumes that the perceived depth values are such that the amount of perceived nonrigidity is minimized. This allows the scheme to deal robustly with deviations from rigidity, and also makes for a very simple computational scheme.

Briefly, this computation operates as follows. At any given point in time  $t$  (with its corresponding image frame), the input to the computation consists of the known  $x_i^t$  and  $y_i^t$  image positions for each point  $i$ . The depth values  $z_i^t$  are unknown. At each time  $t$ , an estimated depth  $\hat{z}_i^t$  is computed for each point. (Note that a coordinate system is used throughout where the  $x$  and  $y$  axes are the horizontal and vertical, respectively, and lie in the image plane. The  $z$  axis is the depth axis; positive values lie toward the observer.)

The relevant data used by the computation are the estimated interpoint distances  $\hat{d}_{ij}^t$  between all pairs of points  $i$  and  $j$ ,

$$\hat{d}_{ij}^t = \sqrt{(x_i^t - x_j^t)^2 + (y_i^t - y_j^t)^2 + (\hat{z}_i^t - \hat{z}_j^t)^2}. \quad (1)$$

The computation starts with an assumed shape estimate for frame 0 where all  $\hat{z}_i^0 = 0$ . In other words, given no initial cues to depth, the object is assumed to be flat in the image plane. Then, given the estimated object shape for time  $t$  as codified in the set  $\{\hat{d}_{ij}^t\}$  of interpoint distances, and the subsequent frame for time  $t+1$  (i.e. the measured image values for  $x_i^{t+1}$  and  $y_i^{t+1}$ ), a set of estimated depth values  $\{\hat{z}_i^{t+1}\}$  are chosen which minimize the amount of nonrigidity in the estimated shape between these two frames. The nonrigidity metric computes a weighted total of the amount of stretch that each 3-dimensional interpoint distance undergoes:

$$\text{amount of nonrigidity} = \sum_{i \neq j} \frac{(\hat{d}_{ij}^{t+1} - \hat{d}_{ij}^t)^2}{(\hat{d}_{ij}^t)^3}. \quad (2)$$

The numerator is the square of the change in interpoint distance. The denominator ensures that a given percent change in interpoint distance is weighted more heavily if the two points are close (with respect to the previous frame's estimated shape). Minimization is carried out using an algorithm of Davidon<sup>9</sup>.

A physical analogue of this model described by Ullman is shown in Fig. 2. The minimization can be carried out as energy minimization in a physical system. After a particular image frame, the internal representation of the object consists of the known  $x_i^t$  and  $y_i^t$  values, and the estimated  $\hat{z}_i^t$  values. At the end of the frame, a construction of rods and springs is made. The rods are positioned at the new known image plane positions. The springs are attached to the rods at the estimated depths, and are at their resting lengths. When the next stimulus frame arrives, the rods are moved in the image plane to their new positions, and the springs ride up and down the rods in order to achieve a minimal energy configuration ("Minimal stretch"). The spring constants are set so as to mimic Eq. 2. The new vertical positions attained by the springs constitute the estimated depth values  $\hat{z}_i^{t+1}$ .

-----  
 Insert Figure 2 about here  
 -----

In this paper, a model is described for the computation of depth from changing relative position cues in multi-dot displays. The model is described as a *relaxation labeling process* (or RLP<sup>10</sup>), which is a local cooperative-competitive network model. The structure of the model treats the separate dots as *objects* and the possible depth values for each dot as a set of *labels* which might be applied to that object. The model involves a process which uses constraints derived from the image data and labelings from previous frames to converge iteratively on a choice of label (i.e. estimated depth value) for each object (i.e. dot).

The paper proceeds as follows. First we describe the model in detail. Next, several simulations of the model are discussed. The model is then contrasted with the Ullman model just described<sup>8</sup>. This leads to a discussion of extensions of the model, in particular to the problem of combining different types of cues to depth.

## 2. A Parallel KDE Model

We now describe a model of the KDE for multi-dot displays, cast as a relaxation labeling process. We first describe relaxation labeling in general. Then, the RLP model for kinetic depth is outlined. Some immediate assumptions and consequences of this model are discussed.

### 2.1. Relaxation labeling processes

Relaxation labeling processes are a form iterative local computation used to solve a "labeling problem". In such a problem, a finite set of objects is given, and for each object, one of a finite set of labels is to be chosen for that object. The process maintains a state vector which codifies the estimated likelihoods of the possible labels at each object. A set of constraints on the state vector results in a support vector, which evaluates the current evidence for each label at each object. The state and support vectors are then combined with the current state using an "update rule", resulting in a new state<sup>11</sup>. This process is iterated as many times as desired, hopefully converging to a high state value for a single label at each object, effectively choosing a mutually consistent labeling of the set of objects.

More formally, in RLP time takes on discrete values,  $t = 0, 1, \dots$ , corresponding to the iterations of the relaxation process. There are  $n$  objects,  $1, \dots, i, \dots, n$ , and the possible labels range over the finite set  $Z$ . A particular label for object  $i$  is designated  $z_i$ . At time  $t$ , the state of the process is summarized in the state vector  $f^t$ . The constraints on labels result in a support vector  $s^t$ , representing the current degree of support for each label at each object. Finally, an update rule results in a new state  $f^{t+1} = F(f^t, s^t)$ . This process is iterated, resulting in a sequence of states  $f^0, f^1, f^2, \dots$ .

At time  $t$ , the state of the process  $f^t$  is a vector of probability distributions  $(f_0^t, f_1^t, \dots, f_i^t, \dots, f_n^t)$ , one for each object, satisfying:

$$\begin{aligned} f_i^t: Z &\rightarrow \mathbf{R} \\ f_i^t(z_i) &> 0 \text{ for all } z_i \in Z \\ \sum_{z_i \in Z} f_i^t(z_i) &= 1. \end{aligned} \quad (3)$$

At any given time  $t$ , the state of a relaxation labeling process represents the current estimates of the relative likelihoods of the possible labels at each object.

The support calculation also results in a vector  $s^t = (s_0^t, s_1^t, \dots, s_i^t, \dots, s_n^t)$  satisfying:  $s_i^t: Z \rightarrow \mathbf{R}$ . In most relaxation labeling process models, the support for a given label  $z_i$  at object  $i$  is a linear sum of the support of all labels at all objects for this particular label:

$$s_i^t(z_i) = \sum_{j=1}^n \sum_{z_j \in Z} c_{ij}^t(z_i, z_j) f_j^t(z_j). \quad (4)$$

$c_{ij}^t(z_i, z_j)$  is the "compatibility coefficient". The value of  $c_{ij}^t(z_i, z_j)$  indicates the extent to which the label  $z_j$  at object  $j$  is compatible with the label  $z_i$  at object  $i$ . This value is weighted by  $f_j^t(z_j)$ , so that a label at an object which has a low state value will contribute little to the support calculation. A large value of  $s_i^t(z_i)$  indicates that label  $z_i$  is compatible with the other object labelings.

The state of the process  $f^t$  and support  $s^t$  are combined using an *update rule*,  $F$ , which mixes the current confidences  $f^t$  with the new evidence  $s^t$ . There is a wide class of models for combining sources of evidence in circumstances such as this<sup>11</sup>. For now, we describe the update rule for relaxation labeling processes originally discussed by Rosenfeld et al<sup>10</sup>:

$$\begin{aligned} F: (f^t, s^t) &\rightarrow f^{t+1} \\ f_i^{t+1}(z_i) &= \frac{f_i^t(z_i) (1 + s_i^t(z_i))}{\sum_{z_i' \in Z} f_i^t(z_i') (1 + s_i^t(z_i'))}. \end{aligned} \quad (5)$$

This formula, although admittedly ad hoc, still has the basic features desired of an update rule. In particular, the larger the support value (relative to the support values for the other labels at an object), the larger the increase in confidence. The denominator is a normalization which ensures that the resulting values of  $f_i^t$  still form a probability distribution function. The form of the function requires constraints to be formed so that  $s_i^t(z_i) > -1$ ; otherwise  $f_i^{t+1}(z_i)$  would become negative, contradicting our assumption.

## 2.2. An RLP model for the KDE

The application of RLP to the kinetic depth problem for multi-point displays is relatively straightforward. The set of objects are the points in the display. The label set is a fixed set of potential depth values for each point. The labeling problem is to choose the appropriate label, or depth value, for each object, or image point. The support function evaluates the support of the image data and current depth estimates

for each possible depth value at each object. The constraints used to generate the support vector are based upon whatever image cues one wishes to include in the model, such as consistency with the previous interpoint distance as in Ullman<sup>8</sup>. The only departures from standard RLP as described above are the addition of gain controls in the support calculation, and the allowance for constraints which change with the appearance of each new stimulus frame.

The KDE stimulus is assumed to be a multi-dot display consisting of a sequence of discrete frames, which appear at times  $t_0, t_1, \dots, t_n$ . For a given time step  $t$ , we will occasionally need to refer to the time, say  $t_k$ , at which the current stimulus frame appeared,  $T_t = \max_{k=0,1,2,\dots} \{t_k \mid t_k \leq t\}$ . Each input frame consists of  $n$  points, called *objects* in relaxation labeling terminology. For any given object  $i$ ,  $1 \leq i \leq n$ , the input data at time  $t$  are the image plane coordinates of that point,  $x_i^t$  and  $y_i^t$ . The output of the process consists in the estimated depth values  $\hat{z}_i^t$ , which are chosen from a finite set of potential depth values,  $Z$ , the label set.

Given these definitions, the RLP model for the kinetic depth effect is given by the following algorithm:

- 1) **Initialize.** Set  $f_i^0(z_i) = \frac{1}{|Z|}$ , where  $|Z|$  is the number of elements in the set  $Z$  (the number of distinct depth labels), and set  $\hat{z}_i^0 = 0$  for all objects  $i$ ,  $1 \leq i \leq n$ , and labels  $z_i \in Z$ .
- 2) **Wait for second frame.** Set  $f_i^t(z_i) = f_i^0(z_i)$  and  $\hat{z}_i^t = \hat{z}_i^0$  for all objects  $i$ ,  $1 \leq i \leq n$ , labels  $z_i \in Z$ , and times  $1 \leq t \leq t_1$ .
- 3) **Iterate.** For  $t = t_1, t_1+1, t_1+2, \dots$ , perform Steps 4 through 6.

4) **Compute  $s^t$ .** For all objects  $i$  and labels  $z_i$ , set

$$s_i^t(z_i) = \alpha g_i^t(z_i) \sum_{j=1}^n \left\{ h_{ij}^t \sum_{z_j \in Z} \left[ c_{ij}^t(z_i, z_j) f_j^t(z_j) \right] \right\}, \quad (6)$$

where

$$c_{ij}^t(z_i, z_j) = G(\Delta \hat{d}_{ij}^t(z_i, z_j), \sigma_{\Delta d}), \quad (7)$$

$$\Delta \hat{d}_{ij}^t(z_i, z_j) = \sqrt{(x_i^t - x_j^t)^2 + (y_i^t - y_j^t)^2 + (z_i - z_j)^2} - \hat{d}_{ij}^{T^t}, \quad (8)$$

$$\hat{d}_{ij}^t = \sqrt{(x_i^t - x_j^t)^2 + (y_i^t - y_j^t)^2 + (\hat{z}_i^t - \hat{z}_j^t)^2} \quad (9)$$

$$G(x, \sigma) = \frac{e^{-x^2/2\sigma^2}}{\sigma\sqrt{2\pi}}, \quad (10)$$

$$h_{ij}^t = G(l_{ij}^t, \sigma_l), \quad (11)$$

$$l_{ij}^t = \sqrt{(x_i^t - x_j^t)^2 + (y_i^t - y_j^t)^2}, \quad (12)$$

$$g_i^t(z_i) = G(\Delta z_i^t(z_i), \sigma_{\Delta z}), \quad (13)$$

$$\Delta z_i^t(z_i) = z_i - \hat{z}_i^{T^t}, \quad (14)$$

and  $\alpha$ ,  $\sigma_{\Delta z}$ ,  $\sigma_l$ , and  $\sigma_{\Delta d}$  are constants.

5) **Compute  $f^{t+1}$ .** For all objects  $i$  and labels  $z_i$ , set

$$f_i^{t+1}(z_i) = \frac{f_i^t(z_i) (1 + s_i^t(z_i))}{\sum_{z_i' \in Z} f_i^t(z_i') (1 + s_i^t(z_i'))}. \quad (15)$$

6) **Compute  $\hat{z}_i^{t+1}$ .** For all objects  $i$ , set

$$\hat{z}_i^{t+1} = z_i \in Z \text{ such that } f_i^{t+1}(\hat{z}_i^{t+1}) = \max_{z_i \in Z} f_i^{t+1}(z_i). \quad (16)$$

Algorithm Steps 1 and 2 are basically initialization. Step 3 controls the iterations. Step 4 and 5 are the basic RLP steps of support calculation and update rule. Finally, Step 6 estimates the depth values from the new state vector computed in Step 5.

The process attempts to compute an estimated depth value for each object. It begins with no knowledge of the various depth values, as represented by a uniform distribution over the  $z_i$  values at each object, as computed in Step 1. Given no knowledge of the possible depth, the process defaults to considering the object as flat.

with all  $\hat{z}_i^0$  equal to zero.

During the entirety of input stimulus frame 0 there is no previous input stimulus or estimated shape with which to compare the current input, and so  $f_i^t$  remains flat, and the process is in a state of ignorance as to depth values. All  $\hat{z}_i^t$  are 0 throughout this period (by default). This is the purpose of Step 2. When a new stimulus frame appears, the relaxation process can begin.

For each time step that a stimulus frame remains displayed, an iteration of the relaxation process takes place (Step 3). The state of the process  $f^t$ , is already available as the result of time step  $t-1$  (from either Step 2 or Step 5).

The support values  $s^t$  are computed in Step 4 using Eq. 6. This step is the heart of the process. As in other RLP models, support for a given label at a given object is computed as a sum of constraints from other labels at other objects, by weighting a compatibility coefficient  $c_{ij}^t(z_i, z_j)$  by the state value  $f_j^t(z_j)$ . In a generalization of RLP, we add gain controls to the calculation:  $h_{ij}^t$  and  $g_i^t(z_i)$ . Finally, the supports are scaled by the term  $\alpha$ , which acts as a rate parameter for the relaxation process.

The compatibility coefficients are computed using  $\Delta \hat{d}_{ij}^t(z_i, z_j)$ , which provides a measure of the change in interpoint distance between objects  $i$  and  $j$  that this pair of  $z$  values would entail, as compared to the estimated interpoint distance value from the end of the final relaxation iteration of the previous stimulus frame. This previous estimate is actually computed just after that time step (in Step 6), and hence the comparison with  $\hat{d}_{ij}^T$ , the estimated interpoint distance available at the beginning of the current stimulus frame.  $G(x, \sigma)$  is the value of the 0-mean  $\sigma$ -standard deviation Gaussian density function evaluated at the point  $x$ . Thus,  $c_{ij}^t(z_i, z_j)$  has a value which is greatest if  $\Delta \hat{d}_{ij}^t(z_i, z_j)$  is zero. Recalling that the compatibility  $c_{ij}^t(z_i, z_j)$  expresses the support that the depth value  $z_j$  at object  $j$  has for the depth value  $z_i$  at object  $i$ , we see support is highest if the values of  $z_i$  and  $z_j$  determine a 3-D Euclidean interpoint distance between objects  $i$  and  $j$  that is equal to their estimated interpoint distance at the end of the preceding stimulus frame. As in Ullman<sup>8</sup>, the process rewards small incremental nonrigidity as measured by interpoint distances.  $\sigma_{\Delta d}$  is a parameter which controls the tolerance for small amounts of nonrigidity.

In a generalization of the usual neighborhood structure of relaxation labeling problems, we include a gain control  $h_{ij}^t$ , which allows us to couple more tightly certain object pairs.  $l_{ij}^t$  is the interpoint distance in the image plane between objects  $i$  and  $j$  at time  $t$ . Examining Eq. 11, we see that constraints are most effective between points that are close in the image. The degree of this unequal gain is controlled by the parameter  $\sigma_l$ .

The constraints are also gated by the term  $g_i^t(z_i)$ . This term allows the inclusion of bias for particular  $z_i$  values, and in this case it allows for a preference for smaller changes in  $z_i$  values from frame to frame.  $\Delta z_i^t(z_i)$  is the amount of change in depth value that label  $z_i$  would imply for object  $i$  relative to its final estimated value from the previous frame. Support is thus amplified if the  $z_i$  value does not differ too strongly from its estimated value in the previous frame, as controlled by the parameter  $\sigma_{\Delta z}$ .

In Step 5, the update rule  $F$  is applied, resulting in  $F(f^t, s^t) = f^{t+1}$ , the state at the beginning of the next time step. The estimated depth values for that time step,  $\hat{z}_i^{t+1}$  are computed from the state vector in Step 6. The depth value is chosen for each object which has the largest confidence value. This cycle of support calculation, update, and depth estimation continues for each time step during which the second frame is available. The state values, depth estimation, and support calculation are illustrated in Fig. 3.

-----  
 Insert Figure 3 about here  
 -----

### 2.3. Comments on the model

The depth is estimated from the state vector in the algorithm by choosing the label with the highest confidence value. In the rare case of ties, preference is given to  $z$  values nearer to the horopter (defined to be a depth of zero). This is the rule which implies the depth values of zero corresponding to flat confidence distributions used in Steps 1 and 2. The maximum rule for the computation of  $\hat{z}_i^t$  is not the only way one might estimate depth from the distribution. For example, the values of the distribution at several points near a peak could be used to interpolate an estimate of the peak. In order to interpret the depth estimates, they must be compared to the depth values used to generate the stimulus (prior to projection),  $z_i^t$ . Given the underdetermination of the structure from motion problem, especially under parallel perspective, the values of  $z_i^t$  only reflect how the input for a given simulation was derived, and comparisons of  $z_i^t$  and  $\hat{z}_i^t$  must necessarily take that underdetermination into account.

Notice that we are assuming, as does Ullman<sup>8</sup>, that the *correspondence problem* for these multi-dot frame sequences has already been solved, since the model is given the sequence of image plane coordinates of an identified object, rather than having to determine which object in a previous frame corresponds to any particular object in the current frame. One might assume a previous correspondence process such as that of Ullman<sup>12</sup>, but this still remains a very large assumption, and the robustness of the model under errors of the given correspondences is an important and still untested issue.

The relaxation process can compute the depth values for frame  $i$  from the time of its appearance at time  $t_k$  until the appearance of the next frame at time  $t_{k+1}$ . If one assumes that relaxation iterations take a fixed amount of time, this may have consequences. If the relaxation process takes several iterations to converge to a correct solution, the model may then predict the outcome of speeding up the motion of a KDE stimulus as simulated by allowing fewer relaxation iterations per input stimulus frame.

The compatibility coefficients,  $c_{ij}^t(z_i, z_j)$  allow inclusion in the model of specific cues derived from the image. The version of the model described by the above algorithm uses a support calculation very similar to the incremental rigidity error metric used by Ullman<sup>8</sup>. On the other hand, there is no reason not to use the compatibilities to utilize

other cues to depth, simply by adding them in as additional constraints. In later sections, the addition of other cues to the compatibilities will be discussed.

The algorithm used for most of the simulations in this paper is actually slightly different than that described above. As described so far, the state vectors are always carried forward from one time step to the next. This process takes place regardless of whether a new time step involves a new display frame of the stimulus or not. In principle, this would allow the depth values from the previous frame to influence those of the next frame by the carried-over state vector values, in addition to influencing them via the constraint calculation. For the moment, we will in fact reset the state at the beginning of all input stimulus frames to a flat distribution. In effect, Step 4 of the algorithm is replaced by the following.

4-a) **Reset  $f^t$ .** If  $t = T_i$  then

Set  $f_i^0(z_i) = \frac{1}{|Z|}$  for all objects  $i$ ,  $1 \leq i \leq n$ , and labels  $z_i \in Z$ .

4-b) **Compute  $s^t$ .** For all objects  $i$  and labels  $z_i$ , set

$$s_i^t(z_i) = \alpha g_i^t(z_i) \sum_{j=1}^n \left\{ h_{ij}^t \sum_{z_j \in Z} \left[ c_{ij}^t(z_i, z_j) f_j^t(z_j) \right] \right\}. \quad (6)$$

Thus, when stimulus frame  $j$  appears at time step  $t_j$ , we have the state as computed by the last iteration  $f^{t_j}$ . This is used to compute the  $\hat{z}_i^{t_j}$  values as usual. Then the state values are reset to flat distributions before computing the support values  $s^{t_j}$ . We will discuss this resetting of the state values later.

There are five parameters which control the process. The label set for any given object is a fixed set of possible depth values,  $Z$ . At any time  $t$ , the estimate of the depth of object  $i$ ,  $\hat{z}_i^t \in Z$ . In addition to the set  $Z$ , there are four parameters used in the support calculation:  $\alpha$ ,  $\sigma_{\Delta z}$ ,  $\sigma_l$ , and  $\sigma_{\Delta d}$ .

### 3. Simulations of the Model

The model described above was used to simulate the computation of depth from moving dot stimuli (KDE) for a variety of configurations. In this section we present the results of several such simulations in which the number of points, number of relaxation iterations per frame, and form of the update rule are varied.

#### 3.1. Initial simulations

The first simulation of the model we shall discuss involves a simple three point stimulus. These three points were rotated about a vertical axis through the origin. The three points were rotated through two complete revolutions, for 48 frames, at an increment of 15 deg per frame.

While working on this initial simulation, a difficulty with the model was discovered which, as we will see in the next section, gets to the heart of the kind of information that the KDE provides. The model as outlined above operates purely on the assumption of parallel perspective. Given this fact, a particular series of stimulus frames representing rigid motion has an infinite number of possible rigid interpretations, because a reversal of all depth values, or the addition of a constant amount of depth to

all depth values would not change the image under parallel perspective. Thus, in a particular frame, depth .1 at object 1 may support depth .6 at object 2, but depth .2 will support depth .7, .3 supports .8, and so on, all to the same extent. The relaxation process has no input or bias rooting it to a particular interpretation distance.

Precisely the same problem exists in Ullman's model<sup>8</sup>. In replicating the results of Ullman's model, solutions tend to oscillate greatly in absolute distance, showing more about the order in which parameters are adjusted in the particular minimization algorithm chosen, than about the stimulus per se. In the relaxation model, several interpretations can be considered simultaneously, in effect, by keeping a probability distribution over the depth values at each point. This absolute depth ambiguity then prevents a single interpretation from winning out in the competition, and the model operates extremely poorly.

In order to circumvent this problem, we have included a bias to keep the interpretation rooted to the image plane. In particular, to any input stimulus, we insert an extra stimulus point at the origin, which never moves. In terms of the algorithm, there is an  $n$ -plus-1st point which is considered to be at a depth of 0 with total confidence (i.e.  $f_{n+1}^t(0) = 1$ ). This point takes part in the support calculation (Step 4 of the algorithm), but its portion of the state vector  $f$  is never changed in Step 5. This is clearly an ad hoc solution to this problem. An alternative would have been to bias the gain control  $g_i^t(z_i)$  for low depth values.

Later, we will discuss this problem in the context of other cues to depth and the kinds of information they provide. For example, one might consider other cues which do yield an absolute distance estimate, and use this information as an input to the process which roots the solution to a particular depth region. The absolute size of the display is not changing throughout the rotation, and so there is no size cue for recession or approach. This information may also serve as an input to the process causing the solution to remain in a particular depth region.

-----  
Insert Table 1 about here  
-----

With the addition of the phantom point at the origin, the model performs quite well. The parameters that were used are given in Table 1. The actual depth values computed by the model for several frames are illustrated in Fig. 4-a. The model starts from a flat interpretation, and requires several frames of input to "grow out" to the correct depth values. The error in the prediction as a function of frame number is given in Fig. 4-b. The ordinate is a normalized predictive interpoint distance error given by

$$\begin{aligned} \text{Error}'(t) &= \sum_{i < j} (d_{ij}^t - d_{ij}^1)^2 \\ \text{Error}(t) &= \frac{\text{Error}'(t)}{\text{Error}'(0)} \end{aligned} \quad (17)$$

This error metric is the same as that used by Ullman<sup>8</sup>, and is normalized to unity for the first frame, i.e. for a "flat" interpretation of the first frame. It is based entirely on interpoint distances, and thus reaches a minimal value (of zero) not only if the model produces estimated depths identical to the intended depths, but also for alternative valid interpretations including those with a constant added to all depth values, and those including a mirror reflection about the image plane (a "depth reversal"). The error value, after an initial rise, falls to a relatively low value. The small periodicity in error after the first rotation is probably an artifact of the error metric.

-----  
Insert Figure 4 about here  
-----

### 3.2. Effect of dot numerosity, number of iterations, and focus of constraints

In Fig. 5-a, the normalized interpoint distance error is plotted as a function of the number of dots in the stimulus. Each curve represents a single simulation for a single multi-dot stimulus rotating about a vertical axis. In general, the model appears relatively indifferent to the number of points, as long as there are at least three points in the stimulus. Note that a two point stimulus often yields a poor depth impression in human observers<sup>13</sup>. In Fig. 5-b an example of the model solution for a 6 point stimulus is shown.

-----  
Insert Figure 5 about here  
-----

The effect of degree of rotation per frame is illustrated in Fig. 6. In Fig. 6-a the error is plotted as a function of the frame number. It appears that the algorithm (after an initial error), is much more effective with the stimulus rotated 60 deg per frame. In Fig. 6-b the data are replotted as a function of rotation angle, rather than frame number, and the convergence seems to be a function of the rotation angle, independent of the amount of rotation per frame in this range. Evidently convergence behavior is driven by information, in some generalized sense, rather than merely by additional frames.

In the Ullman model convergence improves with amount of rotation per frame, for amounts less than that used here, and then stops improving in the range from 30 to 60 deg<sup>8,14</sup>. These results are fairly consistent with those presented here. Given the quantization of depth planes used in this model, it is clear that with very small amounts of rotation per frame, the current model would also converge slower, and in the limit would not find any depth at all.

-----  
Insert Figure 6 about here  
-----

All of the stimuli discussed so far were generated using parallel projection, and hence were "rigid" according to the model's internal representation of the world. In Fig. 7, the convergence behavior is shown for a stimulus generated using a fairly large magnitude of polar projection. Since this stimulus is "nonrigid" as far as the model is concerned, the comparison depth values are, in a sense, not something one could expect the model to find. Nevertheless, the model does converge reasonably well for a time, although near the end of the first rotation, the model loses track of the correct depth values (in fact, it recovers during the next rotation).

-----  
Insert Figure 7 about here  
-----

The simulations so far discussed use a value for  $\sigma_l$  which effectively neutralizes the gain control for locality in the image plane. As Ullman<sup>8</sup> mentions, it would be interesting to note whether a network model such as this would succeed when using only local connectivity and constraints, rather than allowing every stimulus point to interact with all others as we have been doing thus far. One can imagine such a localized interaction as the first step towards a KDE model which can handle multiple objects with separate motion paths. In Fig. 8 the convergence behavior is illustrated with a much smaller value of  $\sigma_l$ , which effectively narrows the focus of the constraints on a particular point to those arising from points nearer in the image (see Table 1;  $\alpha$  was also changed to keep the support values comparable in magnitude). Constraints on more distant points arise only by propagation through intermediate points under these conditions. The convergence behavior is certainly as fast using this narrow focus for  $h_{ij}^t$  as compared to the original wider focus simulation, and in fact appears to be slightly faster. Thus, local constraints can be used effectively in a KDE model.

-----  
Insert Figure 8 about here  
-----

### 3.3. Update rule and convergence behavior

As a model of human kinetic depth performance, the current model is lacking in several directions. One problem is relaxation convergence time. The number of relaxation iterations per stimulus frame has been fixed at 75 for all of the simulations

illustrated thus far. For a network model which is intended to be carried out in a neural substrate, this may well seem like an inordinately large number of iterations<sup>15</sup>. In Fig. 9 we illustrate how the model operates as a function of the number of relaxation iterations per stimulus frame. As the number is decreased from 75 to 25, the model is performing noisily, and with a further reduction to 10 iterations, the model is completely incapable of convergence. (The 180 deg periodic behavior is a function of the 180 deg periodicity of the relative image plane positions of the points, not of the model's depth computation — i.e. it is a side-effect of the particular error metric chosen.)

-----  
 Insert Figure 9 about here  
 -----

The slow convergence problem is, in fact, even worse. As mentioned previously, the original intent was to carry the depth distributions at each point  $f_i^t$  from one frame to the next. This seemed sensible for two reasons. First, this would allow the state information to be much larger over time, allowing prior knowledge of depths to influence further computations (more on this later). Second, it is not clear why there is anything special about the arrival of a new frame which should trigger a new process (the resetting of the  $f_i^t$ 's to be flat distributions). The KDE clearly doesn't suffer given true motion input rather than sampled input; quite the contrary, in fact.

It would therefore be desirable to remove the flattening of the distributions that occurs at the appearance of a new stimulus frame. Unfortunately, as it currently stands, this can not be done without destroying the performance of the model. Several of the above simulations were repeated without the flattening of the distributions (as was the simulation with PLC discussed below). In all cases, the model totally failed to converge. The problem is one of stability. After 75 iterations on frame 2, the distributions have reached a certain strength at the chosen depth values. For the next frame, the computation needs both to flatten that peak and to create a new peak at the new appropriate depth value, if it is to succeed in converging to the correct depth values and thus remain converged. In effect, it requires twice as much change in the  $f_i^t$  values as it did on the previous iteration. At the present time, we have no solution for this problem.

The simulations discussed thus far all utilize the update rule described by Eq. 15. This is only one of many update rules suggested for relaxation labeling, and the field of possibilities grows quite large if this model is viewed in the general context of models for the aggregation of evidence<sup>11</sup>. We have simulated the model using a variety of other update rules, and support calculations. For example, as an attempt to accelerate convergence, we tried raising the support values to various powers,

$$f_i^{t+1}(z_i) = \frac{f_i^t(z_i) (1 + (s_i^t(z_i))^n)}{\sum_{z_i' \in Z} f_i^t(z_i') (1 + (s_i^t(z_i'))^n)} \quad (18)$$

Another version used a support calculation wherein only the depth value with the largest confidence value is allowed to constrain other points and depth values<sup>16</sup>. Finally, we

tried some variations on the Hummel/Zucker relaxation update formula<sup>17</sup>.

In general, these changes of support calculation and update rule had no major effect on the convergence speed and operation of the model. As  $n$  grew larger in Eq. 18, the model did exhibit noisier behavior and a tendency to lose the correct interpretation once having gained it. The same problems occurred occasionally with the Hummel/Zucker operator as the step size parameter of that operator was increased. An example of this is shown in Fig. 10. The error quantity plotted here is normalized depth error, rather than interpoint distance error, as follows:

$$\begin{aligned} \text{Error}'(t) &= \sum_i (z_i^t - \hat{z}_i^t)^2 \\ \text{Error}(t) &= \frac{\text{Error}'(t)}{\text{Error}'(0)}. \end{aligned} \tag{19}$$

The reason for using this new error measure is that it allows us to see that a depth reversal occurred in this particular simulation. With the previous error metric, we would have seen the average interpoint distance error climb at frame 43 and then quickly return again to a low value. Plotted as  $z$  error, we can see that the interpretation was lost and then quickly settled to the reflected interpretation.

-----  
Insert Figure 10 about here  
-----

#### 4. Intermediate Discussion

One important question at this point is how does this model compare with the Ullman incremental rigidity model? Looking into this question raises several issues about the model's performance.

First of all, in one sense the use of a model utilizing relaxation labeling need not result in behavior different from that of the global energy minimization proposed by Ullman<sup>8</sup>. Hummel and Zucker<sup>17</sup> proved that relaxation labeling with symmetric constraints (as we have in Eqs. 7-8) and their update rule results in an algorithm equivalent to a global energy minimization. Thus, with properly chosen constraints, there exists a relaxation labeling model that is precisely equivalent to the incremental rigidity model. Such a model would constitute a different choice of minimizing algorithm, and that is all. Hummel and Zucker<sup>17</sup> also prove convergence of the relaxation algorithm, which means that on any given frame, a single depth value will eventually be chosen for each point. We have not proven that the model converges to the correct depth values over a sequence of frames. It certainly has done so in the simulations, and in as much as the model is similar to that of Ullman, the convergence and stability results of Hildreth and Grzywacz<sup>14</sup> should apply.

On the other hand, the actual support functions we have used (Eq. 6) differ from the energy calculation of Ullman (Eq. 2). A minor problem with Eq. 2 is that the error

measure blows up if any  $d_{ij}^t$  approaches zero. If two stimulus points happen to cross paths in a particular frame, Eq. 2 may cause this pair of points to have an overriding influence on the percept, if their estimated depths are similar. Rather than dividing by  $d_{ij}^t$  (enabling points near each other in 3D distance to have a greater weight in the energy calculation), we use a gain control,  $h_{ij}^t$ , which puts greater weight on pairs of points that are close in the image plane (Eq. 11).

The quantization of the possible depth values makes direct comparisons with Ullman's model difficult on a quantitative level. Clearly, this is an arbitrary feature of the model, and a more realistic implementation would include units tuned to a continuum of overlapping depth ranges. Quantizing depth implies quantization error in the model's output. This is visible both in the noise in the error functions after convergence (e.g. Fig. 4-b), and more importantly, in the higher average error after convergence in this model as compared with that of Ullman.

Ullman refers to the occasional loss of the 3-D structure by his model, and states that occasionally the recovered structure is reversed. Thus, his model suffers occasional "depth reversals", as is also the case with human percepts. There are two remarks to be made here. First, depth reversals in human perception are more complicated than an occasional loss of structure and recomputation. Depth reversals occur quite frequently<sup>18</sup>, and appear to be related to, among other things, eye movements and tracking of particular image features/points<sup>19</sup>. These aspects of reversals are clearly outside of the scope of the two models.

The fact that reversals do occur in the models is no surprise. Both models consist of an energy measure and a minimization algorithm. In both energy measures, the "correct" and "reversed" percepts are minimal in energy in the steady state — with both measures, the veridical estimate of depth over a pair of frames should result in zero energy, which is clearly minimal. Thus, if zero energy is ever achieved, then a pair of shape estimates with no changes in interpoint distances has been found, and this is almost certainly the veridical one (up to changes in absolute depth and reversal). Any loss of this perfect estimate in later frames, including subsequent reversals following loss of structure, is clearly a function of errors made by the minimization algorithm and of the energy surface that the energy measure defines. A more robust function minimizer<sup>20</sup> need never suffer losses of structure.

One final note about contrasts between the relaxation labeling model and that of Ullman. The state space used in relaxation labeling is clearly far larger than that given by the Ullman algorithm. At any given time, the Ullman representation of the stimulus consists of a single depth value for each point. In the RLP model, the state consists of a probability distribution across the possible depths at each point. Although we have been discarding these data across frames (by flattening the distributions), assume for the moment that a future RLP model actually maintains this information. Why would one want a representation of the stimulus that contains this extra information?

When placed in the general context of models for the aggregation of knowledge<sup>11</sup>, it is clear that the RLP paradigm and Ullman's model are only two examples of a larger class of models for combining evidence, where the amount of state information can vary across a wide range. The RLP state provides, for example, a degree of confidence in a

particular depth value once chosen. Thus, the RLP model can differentiate between a flat object in the zero depth plane (a set of distributions all peaked at the same depth value of zero) and total ignorance of the object's structure (a set of flat depth distributions). The Ullman model represents these two situations in an identical manner, and cannot differentiate between the two. It would be interesting to see if biasing the degree of confidence in a current shape estimate can effect the time course to converge on a new structure in the human percept (as has been studied recently by Adelson and Hildreth, pers. comm.).

In this more general context, the relationship between the two models becomes clear. In a sense, Ullman's model tracks the peak of the distributions over depth, and the RLP model tracks the entire distribution over a discrete set of depths. Other possibilities might include, for example, parameterizing the distributions (say, as Gaussians), and tracking the mean and variance (related to the work of Hummel and Landy<sup>21</sup>). This would also provide a means of establishing confidence ratings of the depth estimates — higher confidence would be modeled as lower variance.

In any case, these are interesting questions. They speak to the issue of "representation" of objects in an internal estimation of shape, including the representation of uncertain evidence about this estimate, and how that evidence is combined. We now turn to the combination of evidence about the objects when the sources of evidence include more than one cue.

## 5. Combining Cues

It has become apparent that the problem of combining different cues (e.g. cues to depth) is an important one. Ever since Gibson<sup>22</sup> and others pointed out that there is a multiplicity of cues to depth in the visual environment, there has been much work demonstrating human sensitivity to a wide variety of cues to depth. But, saying that we are sensitive to a particular cue does not answer the question of how the cue is derived from the image and how it is used.

More recently there have been a series of models in computer vision which derive depth from various single cues (stereo, motion, texture, shading, blur, 2D form cues, etc.<sup>23</sup>). These models have had varying success in reconstructing depth from image data, demonstrating that the data may be there, but they are noisy and difficult to obtain. The results of different models for different cues will each obtain evidence which may, in fact, conflict. The problem of combining the outputs of these methods can be seen as critical. The hope is that converging evidence from a variety of cues will result in less noisy estimates of depth than those derived from each cue separately.<sup>24</sup>

### 5.1. Experimental studies of cue combination

It is of interest in this problem of cue combination to probe the mechanism for cue combination in human perception. This can be accomplished by creating stimuli in which two cues are varied independently. This approach has been taken, for example, in the study of shape from texture gradients (with cues such as density, texel shape and orientation)<sup>25</sup>. In KDE, such an approach has also been used<sup>26</sup>.

The results of Doshier et al.<sup>27</sup> are especially relevant for the current work. They investigated the relative contribution of two cues as combined with a kinetic depth stimulus. The basic stimulus was a Necker cube presented with polar perspective, and rotated about a central vertical axis. Such a stimulus (see Fig. 11-a), like all KDE stimuli, can undergo perceptual reversals. Given the polar perspective, the two percepts are either of a rigid cube rotating (say, with the front face moving rightward), and of a highly nonrigid truncated pyramid (rotating with the front face moving leftward). The task was always to designate (e.g. at the stimulus onset) which percept was first seen ("Front-left" or "Front-right").

The image was manipulated in two ways. First, variable amounts of stereo disparity were added favoring one or the other percepts. Second, variable amounts of "Proximity luminance covariance" (PLC) were added. This rather effective cue consists of brightening those edges that are intended to be closer in depth (Fig. 11-a). Again, this cue may be used to favor either of the two percepts, and to varying degree depending on the extent of the luminance difference from front to back. In the experiment, varying degrees of stereo and PLC were added to the basic stimulus, and the two cues were either in agreement or in conflict.

The interest in such an experiment is that it allows one to probe how cues are combined. In Doshier et al.<sup>27</sup>, the data for each individual subject were fit using a simple additive cues model, wherein each level of a given cue adds a certain amount of bias to which percept is chosen, and these levels for each of the two cues are simply added, along with a subject bias term, resulting in a criterion for the rigid percept. The resulting number was compared to an error value (a sample of a standard normal random variable), and if the error value exceeded the criterion, the rigid percept was chosen, otherwise the nonrigid one was chosen. This simple additive cues model was very effective in fitting the data of several subjects, with individual differences appearing in the parameters which estimate the effectiveness of each level of each cue.

The success of this simple model suggests that the combination of cues may be effectively computed in a process model such as the one described here. Evidence in RLP in the form of constraints is combined in an additive fashion (the support calculation), so it seems reasonable to suppose that this model might be extended to add some of the other cues which are often present in KDE displays, such as PLC, stereo disparity, relative motion, occlusion, and so on.

-----  
Insert Figure 11 about here  
-----

## 5.2. Classes of cues to depth

Before discussing in detail how one might go about adding other cues to the model (beyond incremental rigidity/interpoint distance changes), it is of interest to examine the type of information afforded by various cues. Depth cues give information about the

distance from the observer of objects corresponding to image features. There are at least three types of information available depending on the cue: absolute, relative, and ordinal (akin to ratio, interval, and ordinal scales in measurement theory).

Absolute cues give information about the absolute distance from the observer to an object. Stereo disparity can be considered an absolute cue, assuming the viewer knows the eye positions and orientations, as can motion parallax under self-motion (both of these may not be the case<sup>28</sup>). Relative cues give information about the relative distances of objects in depth, but not their absolute distance from the viewer. The failure to provide absolute distance results from an underdetermination of the problem wherein the same image would result either from adding a constant to all depths (in parallel perspective) or by scaling the stimulus and the depths (in polar perspective). Examples of relative cues include the kinetic depth effect (both interpoint distance and relative motion cues), foreshortening, etc. Finally, there is the class of ordinal cues. These cues only specify the order in depth of certain pairs of objects, without constraining the relative distances. Examples of ordinal cues include PLC and occlusion. In addition, there is some evidence for "ambiguous ordinal cues", such as motion shear or texture accretion/deletion, where occlusion is indicated but depth order is ambiguous<sup>29</sup>.

Examining the effects of combining cues across classes is complicated. For example, imagine a multi-dot KDE stimulus with added PLC. If the PLC is consistent with the relative motions, then one would expect the PLC to simply bias the observer towards one of the two reversed interpretations of the object, as was the case in Doshier et al<sup>27</sup>. On the other hand, consider a stimulus in which the PLC is not consistent with either interpretation, for example one in which the brightest dots are those at intermediate depths, and the closest and furthest dots were darker. Here the two cues are in conflict, and it is by no means obvious how this combination would be effected given the two types of information provided by the cues.

Consider how one might reconcile these different types of cues in the current model. Since constraints are already treated additively in the support calculations (Eq. 6), one can simply add other cues into the support. The question is as to how the constraints are to be computed for the various classes of cue. The only cue in the model as described so far is a relative cue. The way it works as a relative cue is clear from the calculation of  $\Delta \hat{d}_{ij}^t(z_i, z_j)$  (Eq. 8), which depends only upon relative depth  $z_i - z_j$ .

With these considerations it becomes clear how other classes of cues might be added to the support calculations. To add an absolute cue, the constraint would support only the absolute depth indicated by the cue. For ordinal cues, the constraint would support all depth values consistent with the indicated ordinal relationship.

Given a model which embodies combinations of different classes of cues, it becomes possible to examine in simulations the effects of cue combination. For example, what happens if stereo information is available for only one dot of a multi-dot KDE stimulus? Does the absolute position of the entire structure follow that point around as its stereo position changes? This might also be tested experimentally by putting a (sparse) KDE stimulus in one eye, and only one matching point in the other eye, although it might be difficult to force a particular correspondence for that point (perhaps by vertical position).

### 5.3. Modeling PLC

In order to begin to test these ideas, we added a constraint corresponding to PLC to the model. PLC is an ordinal constraint; given a point  $j$  which is brighter (and therefore prefers to be closer) than a point  $i$ , depth  $z_j$  at point  $j$  should support all depths  $z_i$  at point  $i$  which are more distant (i.e. all  $z_i$  such that  $z_i < z_j$ , see Fig. 11-b). The change to the model is quite simple. The constraint calculation of Eq. 7 is replaced with

$$c_{ij}^t(z_i, z_j) = G(\Delta d_{ij}^t(z_i, z_j), \sigma_{\Delta d}) + PLC_{ij}^t(z_i, z_j), \quad (20)$$

where  $PLC_{ij}^t(z_i, z_j)$  is defined to be

$$PLC_{ij}^t(z_i, z_j) = \begin{cases} g_{PLC} & \text{if object } i \text{ is brighter than object } j \text{ and } z_i \geq z_j \\ g_{PLC} & \text{if object } j \text{ is brighter than object } i \text{ and } z_i \leq z_j \\ 0 & \text{otherwise,} \end{cases} \quad (21)$$

and  $g_{PLC}$  is a parameter defining the relative strengths of the PLC cue and the rigidity cue. It can be considered a function of the difference in luminance of the two points.

This new model was simulated (with a value of  $g_{PLC} = 0.1$ ), using the same 6 point stimulus as in Fig. 5. The PLC was either used to bias toward the "front moving to the right" (PLC+) or "front moving to the left" (PLC-) percept. As is visible in Fig. 11-c, the model with added PLC works reasonably well, converging slightly faster than the model with no PLC. When plotted as interpoint distance error, the PLC+ and PLC- curves are identical, as is to be expected given the complete symmetry of the situation. On the other hand, when plotted as normalized predicted depth error (as compared with the "front right" interpretation) as in Fig. 11-d, we see that PLC+ has indeed biased the interpretation as it should have, and likewise for PLC-.

It is clear that it is possible to use a process model of depth interpretation such as the one outlined here to investigate cue combination. The addition of PLC to the incremental rigidity cue had precisely the desired effects. In addition, it also led to faster relaxation convergence. When the second frame appears in this stimulus, without PLC only three of the six points develop any nonzero depth at all, and it takes most of the 75 relaxation iterations before the third of these three points develops this depth. On the other hand, given the unambiguous character of the ordinal PLC cue, convergence is far faster. Starting with flat distributions, in one single relaxation iteration all six points are in the correct depth order as indicated by the PLC, which is why the PLC convergence in Fig 11-c starts out so much quicker.

### 6. Discussion

We have described a model for the KDE in the form of a cooperative-competitive network, described in the language of relaxation labeling. The model successfully computes depth values in a manner similar to that of the incremental rigidity model of Ullman<sup>8</sup>, although the precise equations are somewhat different. In addition, we have discussed how such a process model may be extended to investigate cue combination, and have tried out these ideas on the simple case of proximity luminance covariance (PLC).

The model is not currently in a state where it might be used to fit psychophysical data since it is completely deterministic. To remedy this would require inclusion of a source of noise<sup>30</sup>.

The model as it stands only represents one piece of the picture as far as KDE is concerned. The model only concerns tracked dots, or any tracked feature points such as endpoints of lines in vector drawings, or trackable features in natural images. The assumption of known feature correspondences is a major one, and the robustness of the model's computations under errors in correspondence should be investigated and compared with human performance. Also, KDE from tracked features is only one form of KDE, there is also KDE from tracked occluding contours which don't correspond to single positions on an object<sup>31</sup>.

In the context of multi-dot stimuli, the model uses only the cue of changing interpoint distance/incremental rigidity. It ignores other cues such as relative motion of dots (which may be a cue to relative depth, as in the optic flow models, or to ambiguous ordinal depth, similar to motion shear), dot density, and the dynamic foreshortening of groups of dots (i.e. deformation, as used by the model of Koenderink and van Doorn<sup>32</sup>), any one of which may turn out to be more important for the human percept. This is a matter for further study. Finally, any scheme of incremental rigidity should eventually prefer a rigid interpretation of a stimulus if one can be found (although it may become caught in a local minimum in the energy function which does not correspond to this interpretation). On the other hand, in cases of objects extending more in depth than in visible breadth, nonrigid percepts of rigid objects are quite common<sup>33</sup>.

To conclude, we have defined a model for the KDE. The model is by no means a final answer, and we have identified a number of problems with it. On the other hand, it is a real attempt at a process model of the KDE, and shows some promise for being capable of dealing with some of the complexities of the phenomenon. Finally, we have discussed some of the difficulties of modeling cue combination in the context of this KDE model, and have pointed the way towards solutions.

## 7. Acknowledgment

The work described in this paper was supported in part by a grant from the Office of Naval Research, Grant N00014-85-K-0077. We would like to thank George Sperling, who first suggested this line of research and the general approach. We also would like to thank Barbara Doshier, Bob Hummel, Charles Chubb, and Mark Perkins, for many helpful discussions, suggestions, and encouragement.

## Notes

1) See, e.g. Gibson, J. J., *Perception of the Visual World*, Boston: Houghton-Mifflin, 1950; Kaufman, L., *Sight and Mind, An Introduction to Visual Perception*, New York: Oxford University Press, 1974.

2) Braunstein, M. L., Depth perception in rotating dot patterns: Effects of numerosity and perspective, *Journal of Experimental Psychology* **64**, 415-420 (1962). Green, B. F., Jr., Figure coherence in the kinetic depth effect, *Journal of Experimental Psychology* **62**, 272-282 (1961). Wallach, H., and O'Connell, D. N., The kinetic depth effect, *Journal of Experimental Psychology* **45**, 205-217 (1953).

3) Andersen, G. J., and Braunstein, M. L., Dynamic occlusion in the perception of rotation in depth, *Perception & Psychophysics* **34**, 356-362 (1983). Proffitt, D. R., Bertenthal, B. I., and Roberts, R. J., The role of occlusion in reducing multistability in moving point-light displays, *Perception & Psychophysics* **36**, 315-323 (1984). Schwartz, B. J., and Sperling, G., Luminance controls the perceived 3-D structure of dynamic 2-D displays, *Bulletin of the Psychonomic Society* **21**, 456-458 (1983).

4) Andersen, G. J., and Braunstein, M. L., Dynamic occlusion in the perception of rotation in depth, *Perception & Psychophysics* **34**, 356-362 (1983). Braunstein, M. L., Depth perception in rotating dot patterns: Effects of numerosity and perspective, *Journal of Experimental Psychology* **64**, 415-420 (1962). Green, B. F., Jr., Figure coherence in the kinetic depth effect, *Journal of Experimental Psychology* **62**, 272-282 (1961). Landy, M. S., Doshier, B. A., and Sperling, G., Assessing kinetic depth in multi-dot displays, *Bulletin of the Psychonomic Society* **19**, 23 (1985). Lappin, J. S., Doner, J. F., and Kottas, B. L., Minimal conditions for the visual detection of structure and motion in three dimensions, *Science* **209**, 717-719 (1980). Petersik, J. T., Three-dimensional object constancy: Coherence of a simulated rotating sphere in noise, *Perception & Psychophysics* **25**, 328-335 (1979). Petersik, J. T., The effects of spatial and temporal factors on the perception of stroboscopic rotation simulations, *Perception* **9**, 271-283 (1980).

5) For a review, see Ullman, S., Recent computational results in the interpretation of structure from motion, In *Human and Machine Vision* (Eds. A. Rosenfeld, B. Hope, and J. Beck), New York: Academic Press, 1983, pp. 459-480.

6) Hoffman, D. D., and Flinchbaugh, B. E., The interpretation of biological motion, *Biological Cybernetics* **42**, 195-204 (1982). Tsai, R. Y., and Huang, T. S., Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces, University of Illinois at Urbana-Champaign Coordinated Science Laboratory Report R-921, 1982. Ullman, S., *The Interpretation of Visual Motion*, Cambridge, MA: MIT Press, 1979.

7) Clocksin, W. F., Perception of surface slant and edge labels from optical flow: a computational approach, *Perception* **9**, 253-269 (1980). Koenderink, J. J., and van Doorn, A. J., Depth and shape from differential perspective in the presence of bending

deformations, *Journal of the Optical Society of America, A* **3**, 242-249 (1986). Longuet-Higgins, H. C., and Prazdny, K., The interpretation of a moving retinal image, *Proceedings of the Royal Society of London, Series B* **208**, 385-397 (1980).

8) Ullman, S., Maximizing rigidity: the incremental recovery of 3-D structure from rigid and nonrigid motion, *Perception* **13**, 255-274 (1984).

9) Davidon, W. C., Variance algorithm for minimization, *The Computer Journal* **10**, 406-413 (1968).

10) Hummel, R. A., and Zucker, S. W., On the foundations of relaxation labeling processes, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-5**, 267-287 (1983). Rosenfeld, A., Hummel, R. A., and Zucker, S. W., Scene labeling by relaxation operations, *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-6**, 420-433 (1976).

11) See, e.g., Landy, M. S., and Hummel, R. A., A brief survey of iterative knowledge aggregation methods, Proceedings of the Eighth International Conference on Pattern Recognition, Paris, France, 1986, In press.

12) Ullman, S., *The Interpretation of Visual Motion*, Cambridge, MA: MIT Press, 1979.

13) Landy, M. S., Doshier, B. A., and Sperling, G., Assessing kinetic depth in multi-dot displays, *Bulletin of the Psychonomic Society* **19**, 23 (1985).

14) Hildreth, E. C., and Grzywacz, N. M., The incremental recovery of structure from motion: Position vs. velocity based formulations, Proceedings of the Workshop on Motion: Representation and Analysis, IEEE Computer Society #696, Charleston, South Carolina, May 7-9, 1986.

15) Ballard, D. H., Cortical connections and parallel processing: Structure and function, *The Behavioral and Brain Sciences* **9**, 67-120 (1986).

16) As in, e.g., Sperling, G., Binocular vision: A physical and neural theory, *The American Journal of Psychology* **83**, 461-534 (1970).

17) Hummel, R. A., and Zucker, S. W., On the foundations of relaxation labeling processes, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-5**, 267-287 (1983).

18) Howard, I. P., An investigation of a satiation process in the reversible perspective of revolving skeletal shapes, *Quarterly Journal of Experimental Psychology* **13**, 19-33 (1961).

19) Peterson, M. A., and Hochberg, J., Opposed-set measurement procedure: A quantitative analysis of the role of local cues and intention in form perception, *Journal of Experimental Psychology: Human Perception and Performance* **9**, 183-193 (1983).

- 20) E.g. "simulated annealing", Geman, S., and Geman, D., Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**, 721-741 (1984); or probabilistic learning automata, Thathachar, M. A. L., and Sastry, P. S., Relaxation labeling with learning automata, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-8**, 256-268 (1986).
- 21) Hummel, R. A., and Landy, M. S., A statistical viewpoint on the theory of evidence, Robotics Report No. 57, New York University, 1986, Submitted for publication.
- 22) Gibson, J. J., Perception of the Visual World, Boston: Houghton-Mifflin, 1950.
- 23) For a brief review, see Nevatia, R., Machine Perception, Englewood Cliffs, NJ: Prentice-Hall, 1982.
- 24) Aloimonos, J., and Rigoutsos, I., Determining the 3-D motion of a rigid planar patch without correspondence. under perspective projection, Proceedings of the Workshop on Motion: Representation and Analysis, IEEE Computer Society #696, Charleston, South Carolina, May 7-9, 1986.
- 25) Cutting, J. E., and Millard, R. T., Three gradients and the perception of flat and curved surfaces, *Journal of Experimental Psychology: General* **113**, 198-216 (1984).
- 26) Schwartz, B. J., and Sperling, G., Luminance controls the perceived 3-D structure of dynamic 2-D displays, *Bulletin of the Psychonomic Society* **21**, 456-458 (1983). Doshier, B. A., Sperling, G., and Wurst, S., Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3D structure, *Vision Research* **26**, 973-990 (1986).
- 27) Doshier, B. A., Sperling, G., and Wurst, S., Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3D structure, *Vision Research* **26**, 973-990 (1986).
- 28) Heinemann, E. G., Tulving, E., and Nachmias, J., The effect of oculomotor adjustments on apparent size, *American Journal of Psychology* **72**, 32-45 (1959). Ono, M. E., Rivest, J., and Ono, H., Depth perception as a function of motion parallax and absolute-distance information, *Journal of Experimental Psychology: Human Perception and Performance* **12**, 331-337 (1986).
- 29) Farber, J. M., and McConkie, A. B., Optical motions as information for unsigned depth, *Journal of Experimental Psychology: Human Perception and Performance* **5**, 494-500 (1979).
- 30) For an example of this in a KDE model, see Inada, V. K., Hildreth, E. C., Grzywacz, N. M., and Adelson, E. H., The perceptual buildup of three-dimensional structure from motion, *Investigative Ophthalmology and Visual Science, Supplement* **27**, 142

31) Todd, J. T., Perception of structure from motion: Is projective correspondence of moving elements a necessary condition?, *Journal of Experimental Psychology: Human Perception and Performance* **11**, 689-710 (1985).

32) Koenderink, J. J., and van Doorn, A. J., Depth and shape from differential perspective in the presence of bending deformations, *Journal of the Optical Society of America, A* **3**, 242-249 (1986).

33) See, e.g., Adelson, A. H., Rigid objects that appear highly non-rigid, *Investigative Ophthalmology and Visual Science, Supplement* **26**, 56 (1985).

### Figure Legends

1) The Kinetic Depth Effect. A transparent cylinder with dots painted on it is rotated in a series of discrete steps. Each position of the cylinder is projected onto an image plane using parallel projection. A single frame has little or no cues to depth, and yet the sequence of frames yields a strong and convincing impression of depth.

2) A physical analogue of the Ullman<sup>8</sup> incremental rigidity model. The rods project out of the image plane at the positions of the dots in a KDE stimulus. The springs are set to be at resting length for the current depth estimates of each point. Given a new frame, the rods are moved to the new image points, and the springs ride up and down the rods in order to achieve a minimal energy configuration. The new endpoints of the springs constitute the new depth estimates.

3) Constraint and support in the relaxation labeling model of the KDE. This is a top view of two points in a KDE stimulus. Next to each point is a line representing the range of possible depth values that may be assigned to the point. The state value is a probability distribution across those depth values, representing relative confidence in each depth. The peak in each distribution is the current estimated depth for that point  $\hat{z}_i^t$ . Each depth at each point can constrain each depth at each other point. The value of the constraint is basically the confidence value weighted by the connecting coefficient. These constraints are summed to form the support for any particular depth at a given point.

4) a) Top view of a three point stimulus, and the depths calculated by the model. The estimate is initially flat (no depth), and slowly grows out to be an accurate estimate of the actual object. b) Convergence behavior for the three point stimulus. The error is the mean square error of the estimated interpoint distances, normalized to 1 for the first frame.

5) a) Effect of the number of points. b) An example of the depths calculated by the model for a 6 point stimulus (this is a top view as in Fig. 4-a).

6) Effect of the rotation angle per frame for a 3 point stimulus. a) Plotted as a function of stimulus frame number. b) Plotted as a function of degrees of rotation.

7) Effect of a "nonrigid" input. Here the stimulus is actually a rigid 6 point stimulus using polar projection. It is nonrigid in the model's terms, since the model assumes parallel projection.

8) Effect of narrow focus for a 20 point stimulus. The focus of the interpoint constraints is narrowed by decreasing  $\sigma_l$  (see Table 1).

9) Effect of number of relaxation iterations per stimulus frame for a three point stimulus. The parameter is the number of iterations per frame.

10) A perceptual "reversal", which occurred using the Hummel/Zucker update rule with a three point stimulus. The error plotted is the mean square error in the estimated  $z$  values as compared with the "correct" values, and as compared with the reflected (sign-reversed) values (both normalized to 1 for the first frame). Around frame 44 the structure was temporarily lost, and the recovered structure was reversed.

11) Effect of adding cues. a) A Necker cube in polar perspective with added PLC (proximity luminance covariance<sup>26</sup>), here represented as thicker lines. The face with the thicker lines is more likely to be perceived as closer to the observer. b) A PLC constraint for the KDE model. If the pair of points and depths are consistent with the brightness cue, then a fixed amount of support,  $g_{PLC}$ , is added. c) The effect of positive and negative PLC on interpoint distance error for a six point stimulus. Convergence is faster with PLC, and PLC+ and PLC- yield identical convergence. d) Plotted as  $z$  error, it is clear that PLC+ created a bias for the "correct" interpretation, and PLC- for the reversed interpretation.

Table 1) Parameter values used in the simulations.

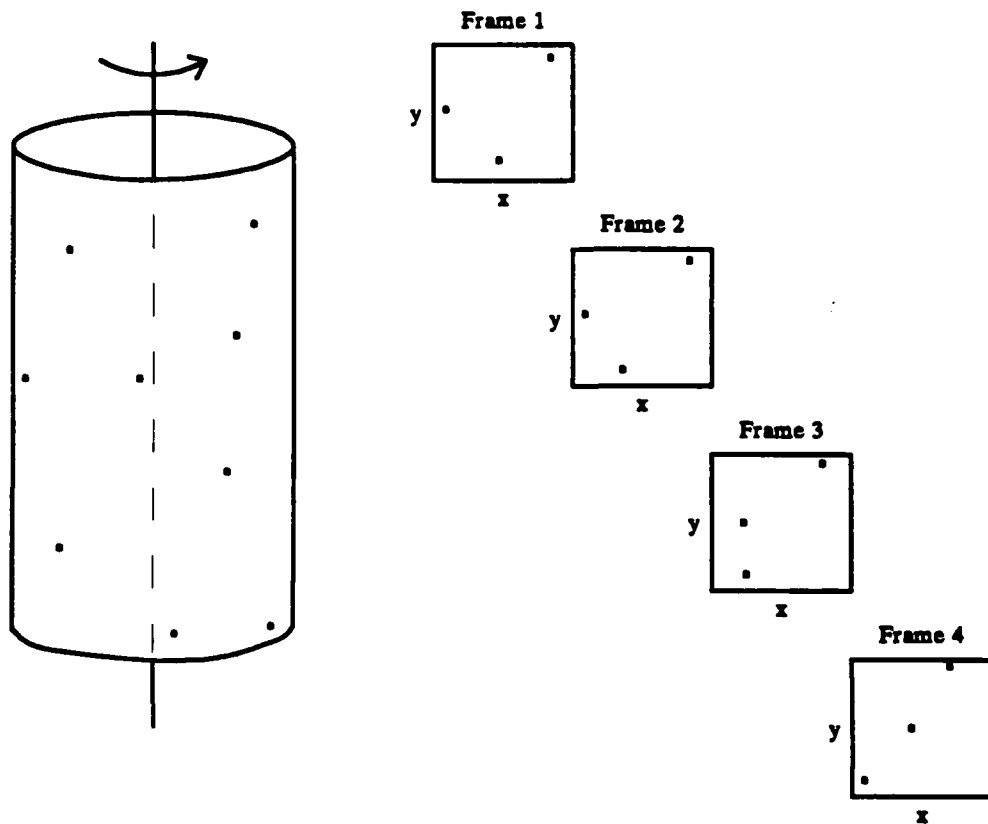


Figure 1

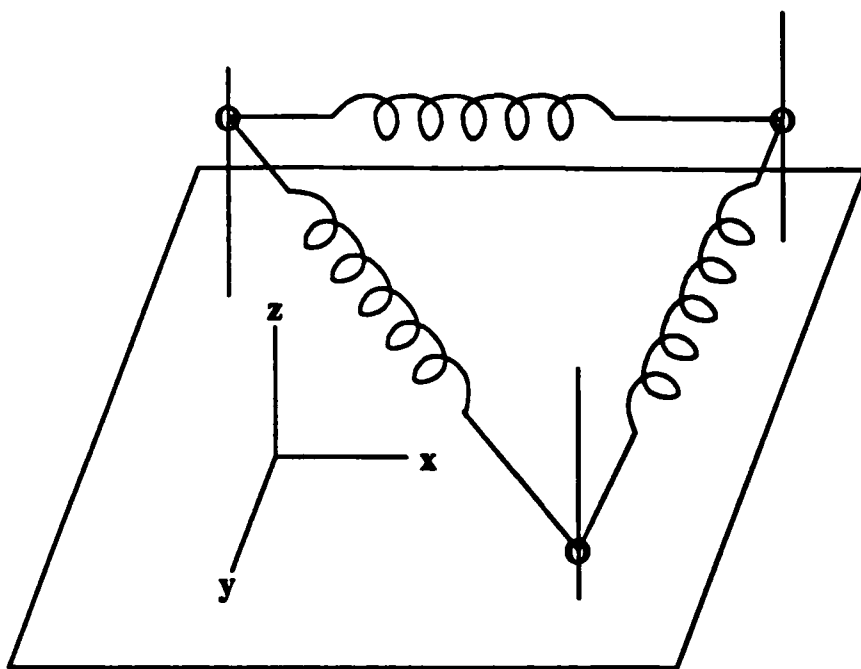


Figure 2

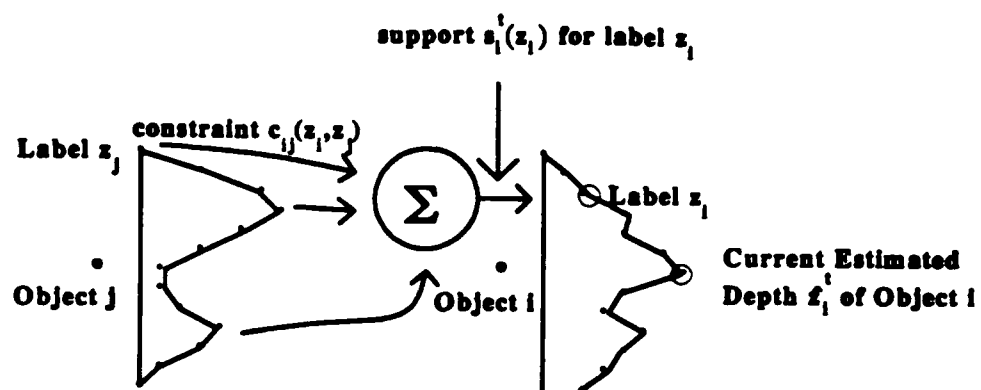
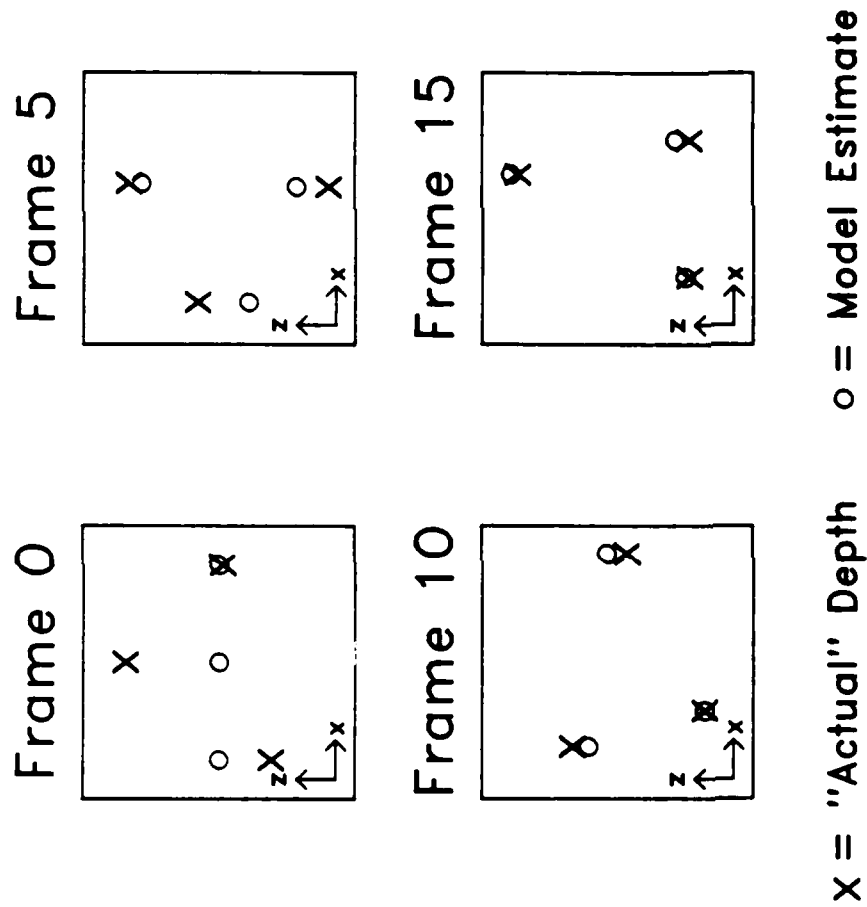


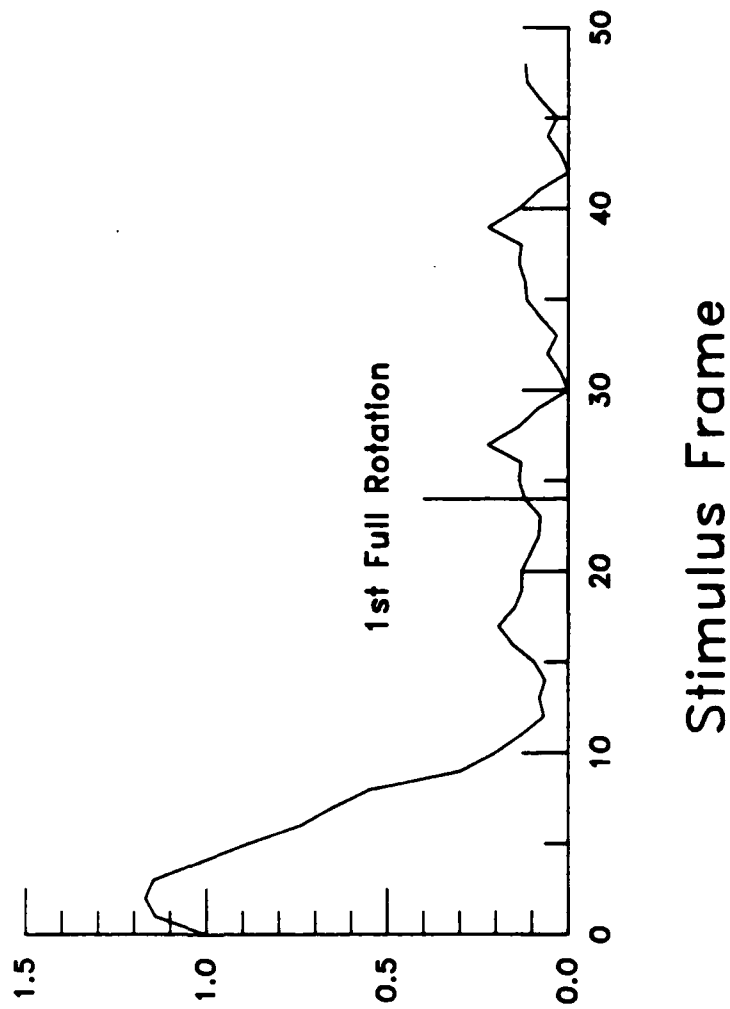
Figure 3

# A

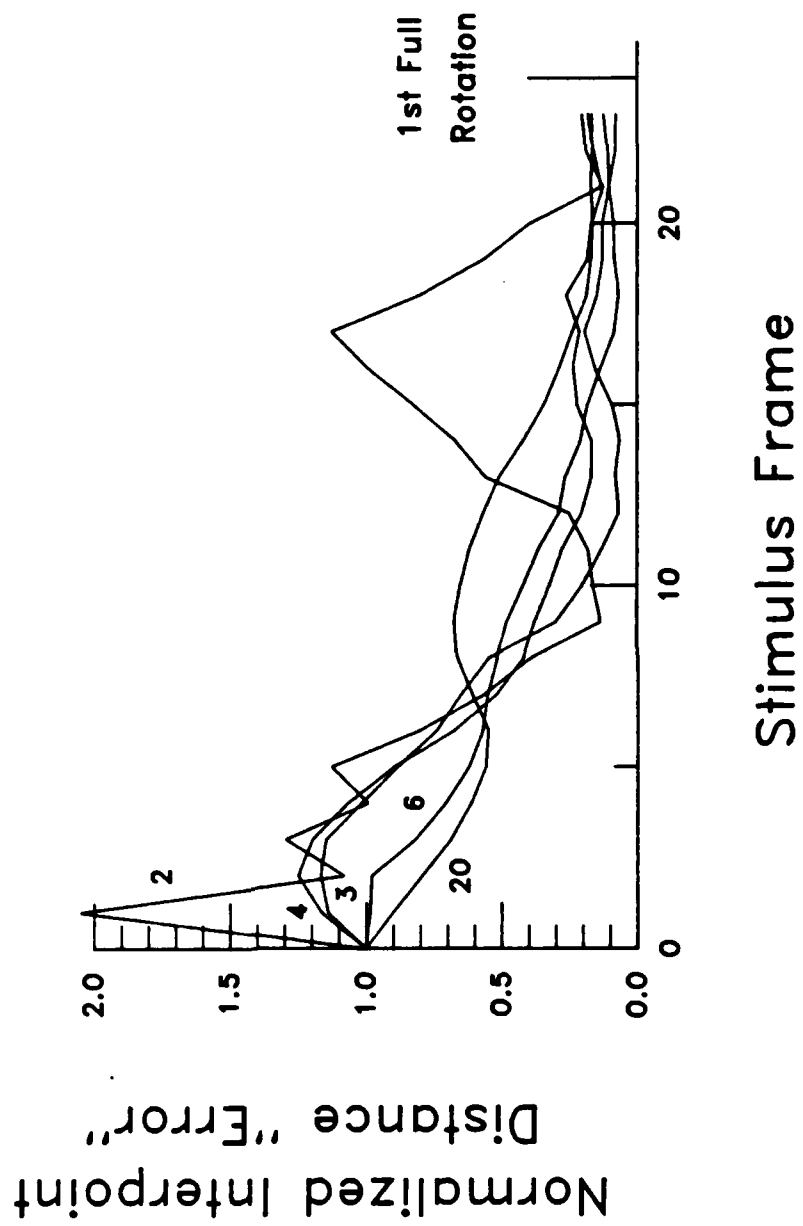


**B**

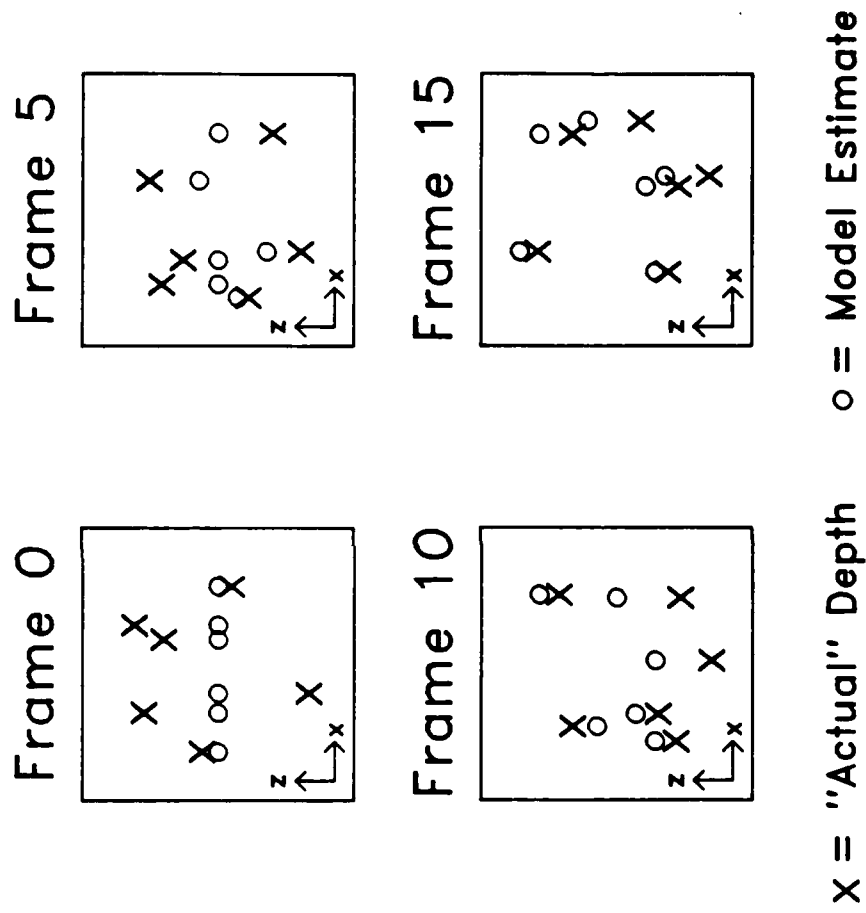
Normalized Interpoint  
Distance "Error"



A

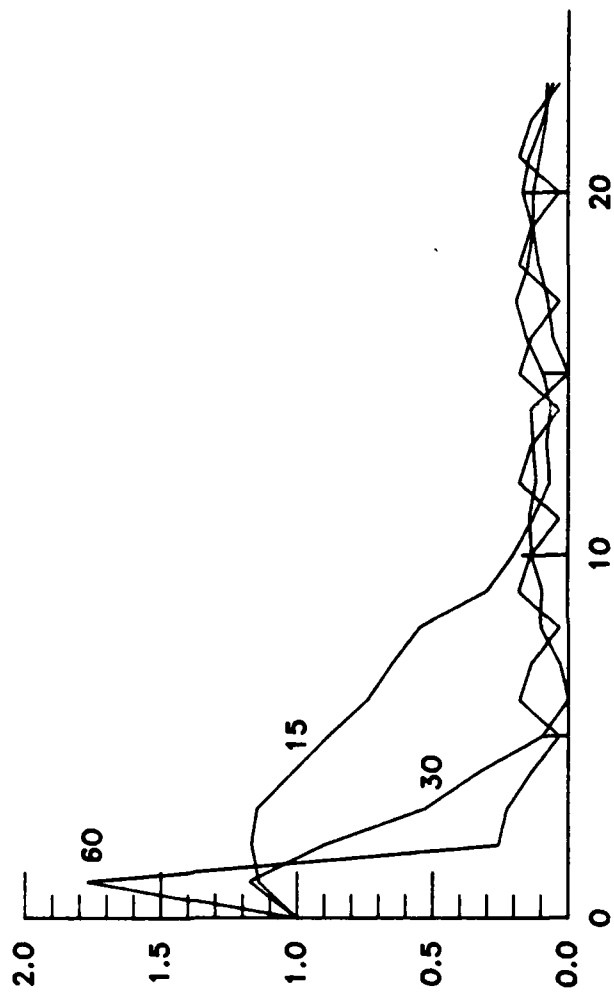


# B



A

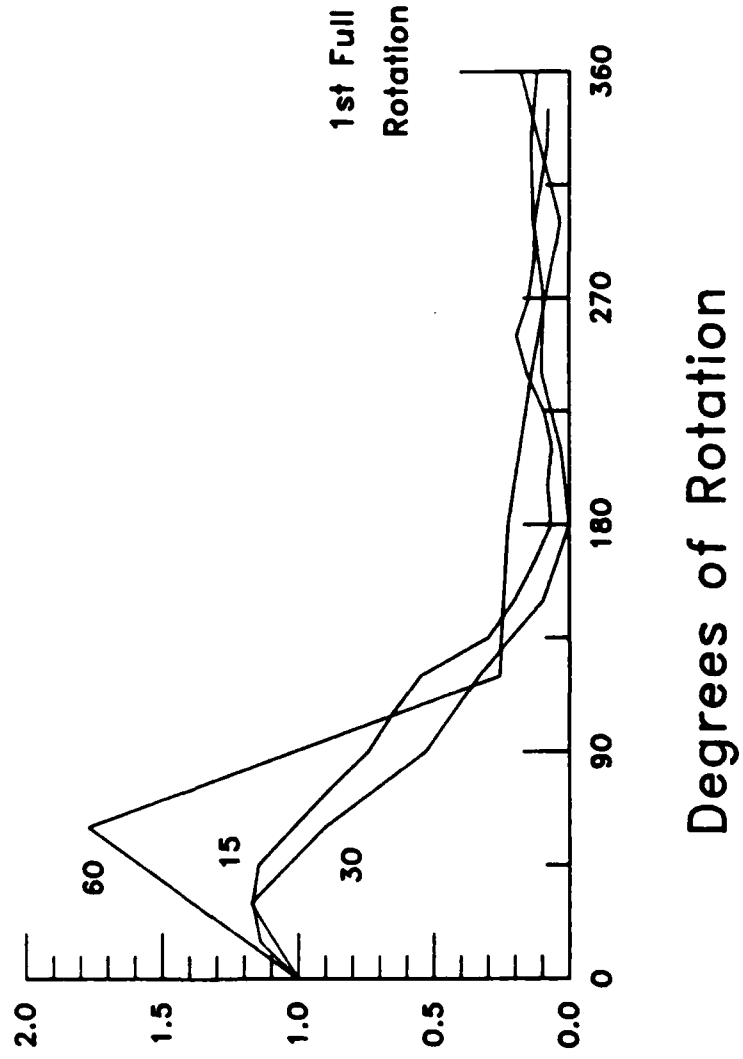
Normalized Interpoint  
Distance "Error"

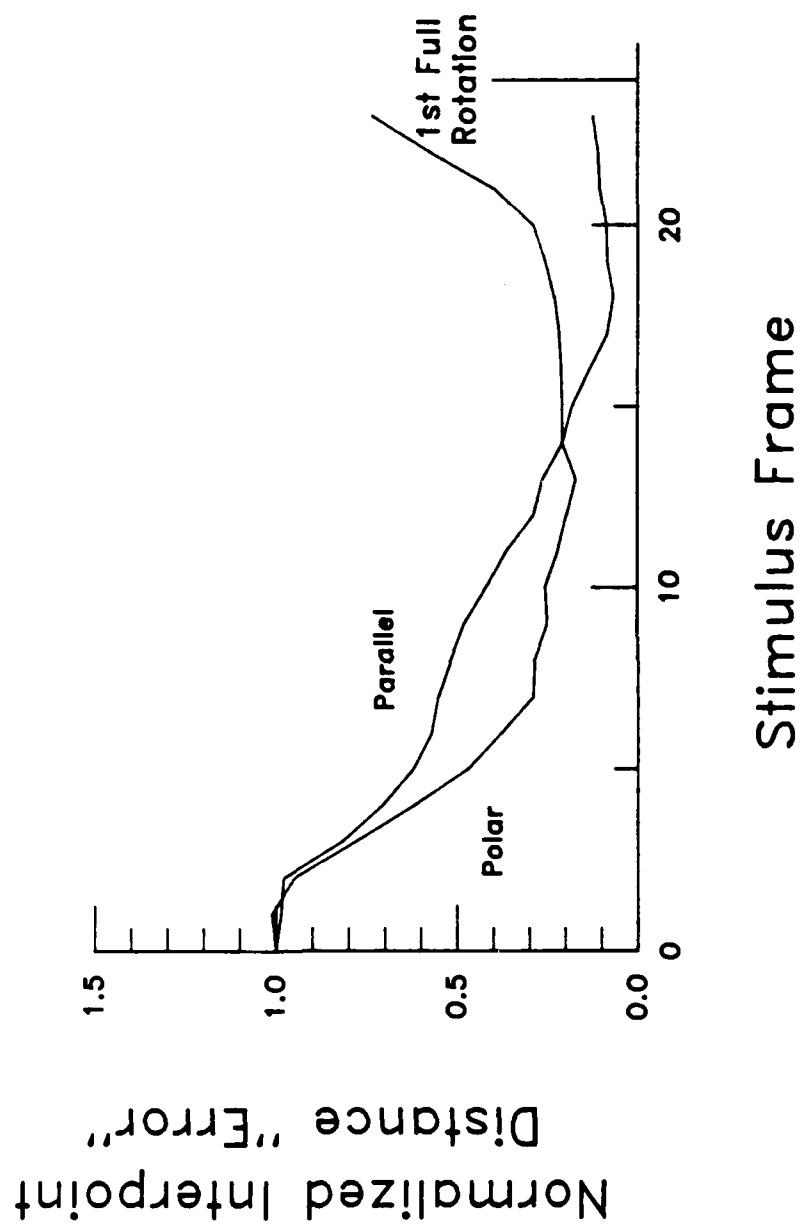


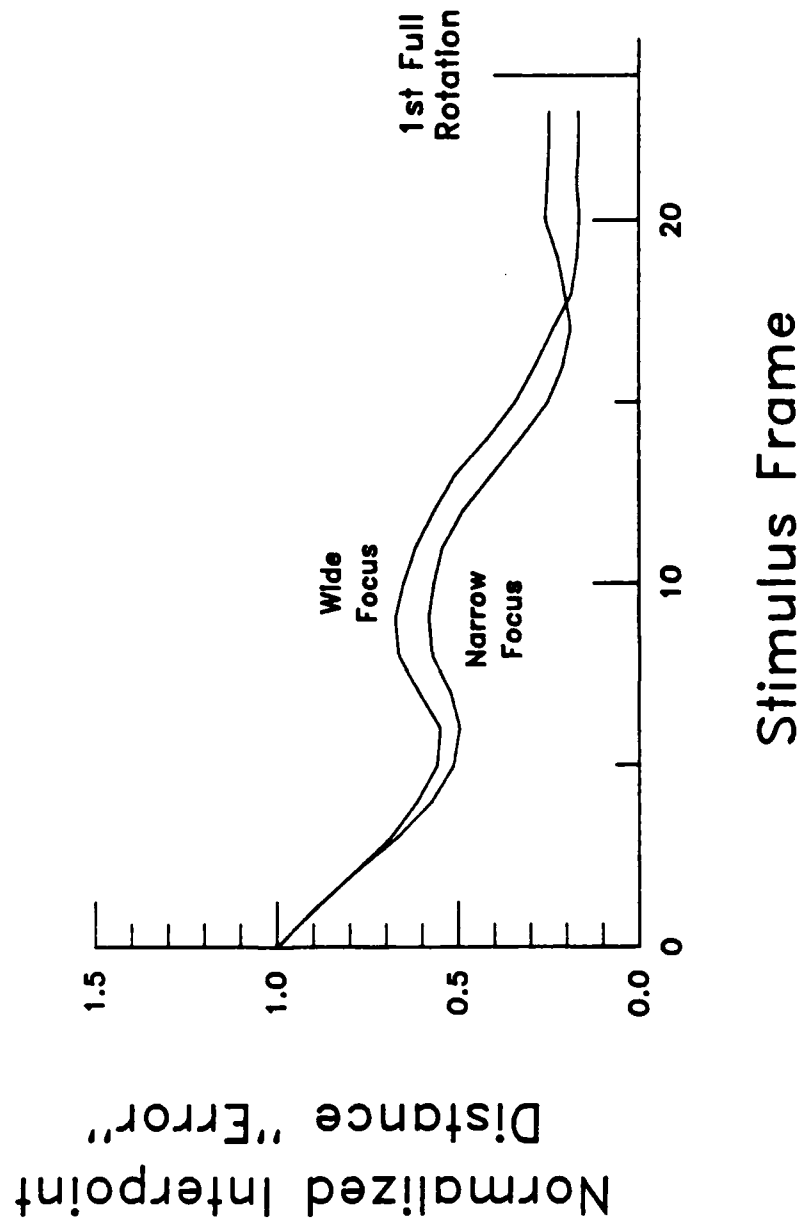
Stimulus Frame

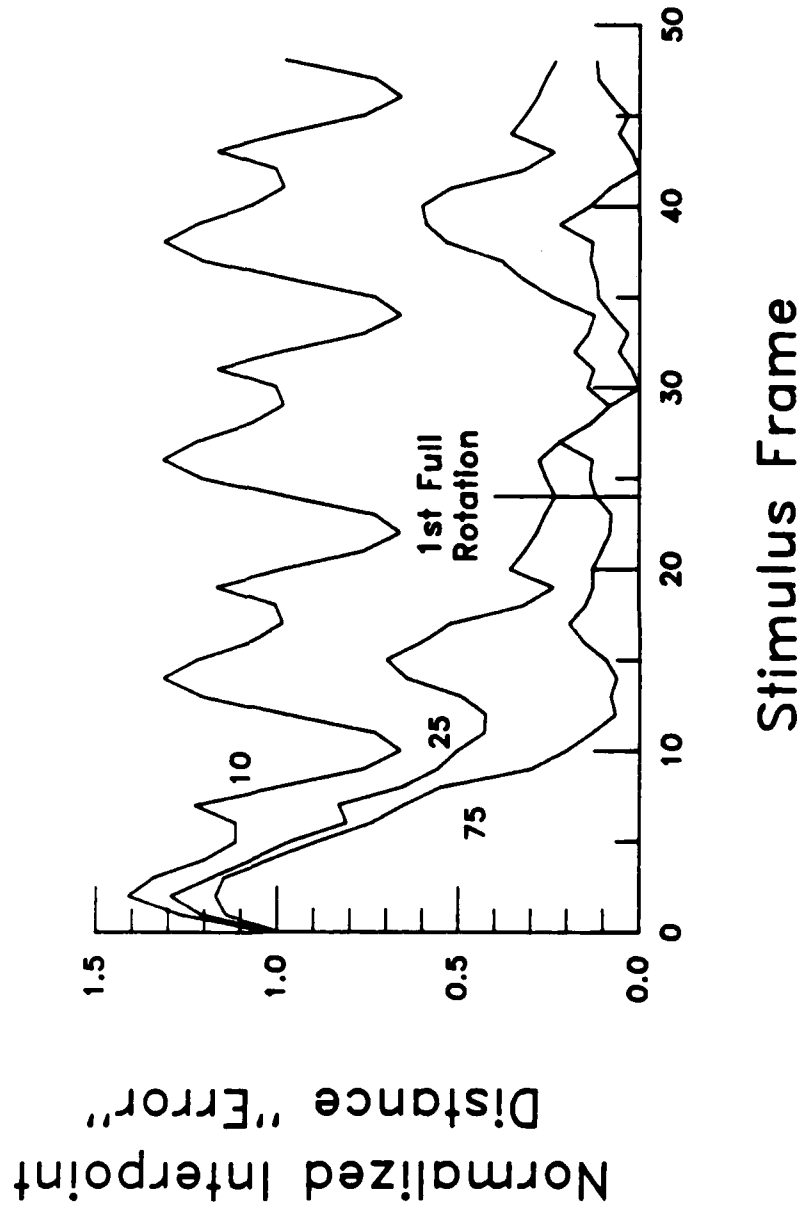
**B**

Normalized Interpoint  
Distance "Error"

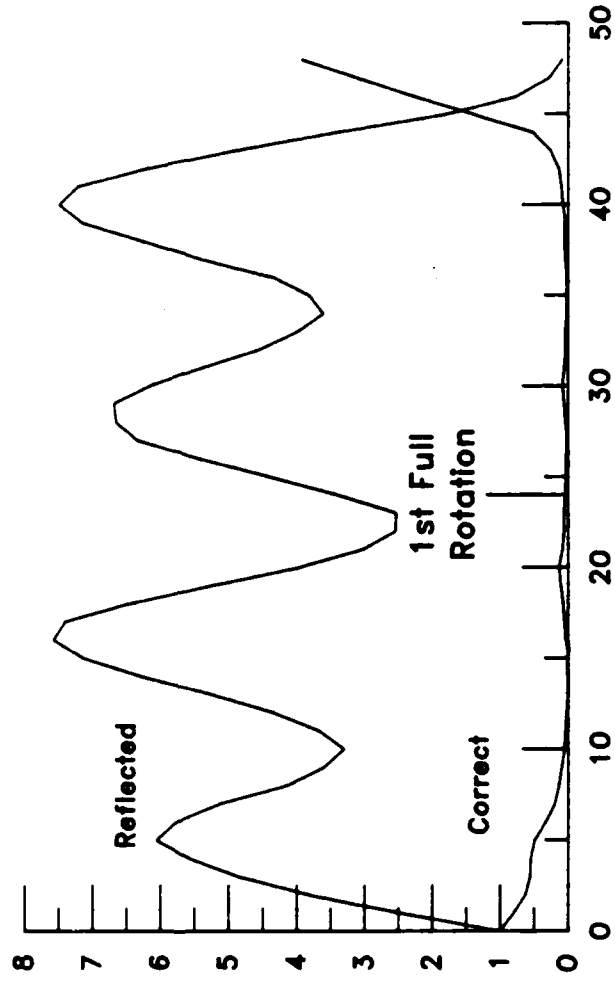






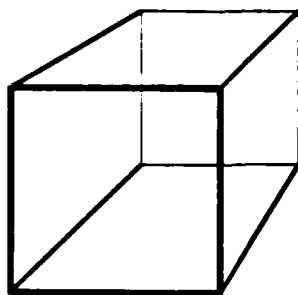


Normalized Predicted  
Z Value "Error"



Stimulus Frame

**A**



**Figure 11-a**

**B**

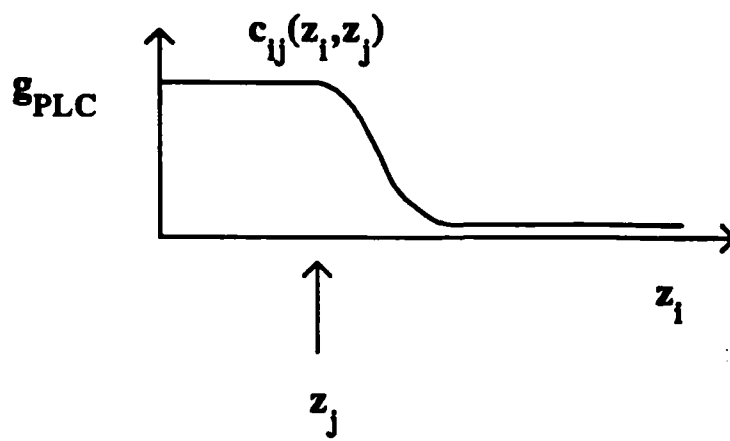
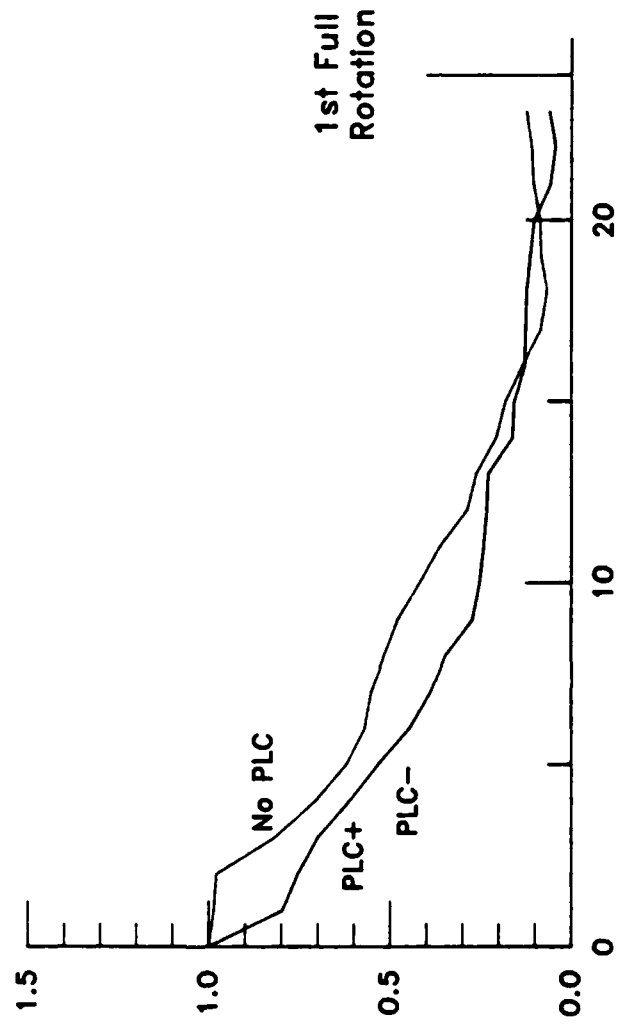


Figure 11-b

C

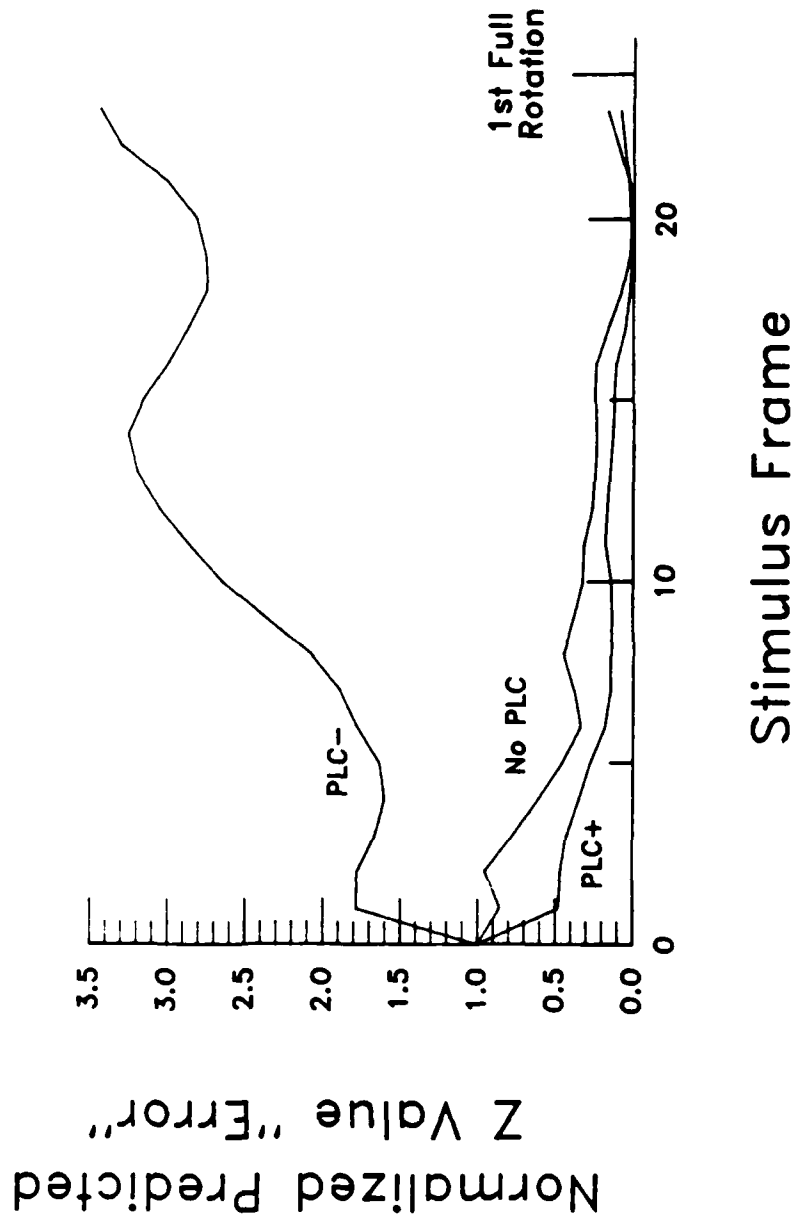
Normalized Interpoint  
Distance "Error"



Stimulus Frame

1st Full  
Rotation

D



Parameter	Meaning	Normal Value	Narrow Focus
$\alpha$	Relaxation step size	30	11
$\sigma_{\Delta z}$	Strength of constraint for small change in depth	4	4
$\sigma_l$	Strength of differential gain for closer points in the image plane ("locality")	3	0.7
$\sigma_{\Delta d}$	Tightness of tuning of constraint for small change in 3-D interpoint distances	0.3	0.3
$Z$	Set of possible depth values	$\{-1.1, -1.0, \dots, 1.0, 1.1\}$	$\{-1.1, -1.0, \dots, 1.0, 1.1\}$
$g_{PLC}$	Strength of PLC constraint	0.1	NA

Table 1

Parameter values used in the simulations.

END

8-87

DTIC